Fair Diffusion: Instructing Text-to-Image Generation Models on Fairness

Year: 2023 | Citations: 167 | Authors: Felix Friedrich, P. Schramowski, Manuel Brack, Lukas Struppek, Dominik Hintersdorf

---

Abstract

Generative AI models have recently achieved astonishing results in quality and are consequently employed in a fast-growing number of applications. However, since they are highly data-driven, relying on billion-sized datasets randomly scraped from the internet, they also suffer from degenerated and biased human behavior, as we demonstrate. In fact, they may even reinforce such biases. To not only uncover but also combat these undesired effects, we present a novel strategy, called Fair Diffusion, to attenuate biases after the deployment of generative text-to-image models. Specifically, we demonstrate shifting a bias, based on human instructions, in any direction yielding arbitrarily new proportions for, e.g., identity groups. As our empirical evaluation demonstrates, this introduced control enables instructing generative image models on fairness, with no data filtering and additional training required.